ICTNET at Web Track 2010 Spam Task

Liang Zhu[1,2], Bolong Zhu[1,2], Jianguo Wang[1,2], Xu Chen[1,2], Zeying Peng[1,2], Xiaoming Yu[1], Yue Liu[1], Hongbo Xu[1], Xueqi Cheng[1]
1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190
2. Graduate School of Chinese Academy of Sciences, Beijing, 100190

Abstract

Web Spamming refers those web pages deceive search engines so as to get a higher rank in their search result. We work on the data set TrecWeb09, based on a content-based spamming classifier, to check the two ends of a hyperlink; if the two end pages either is content spamming, or both are not so good, then the hyperlink will be discarded. After all hyperlinks have been checked, PageRank value shall be re-count on the re-built web network. The balance of one page's PageRank value will be regarded as its link spamming. Then the link spamming score and the result of content deceiving analyzer will be combined as the final estimation of one page's spamming.

1 Introduction

To detect web spamming is becoming an important work of the search engine. How to judge each web page in such a huge world wide web, and to give adequate punishment, is a researchable problem. To deceive the search engine is not the nature property of a web page; it is totally artificial and vicious. We utilize the result of a content based spamming classifier and focus on the final target of web spamming – getting a higher score while search engines evaluate web pages, to do a novel way detecting spamming.

Machine learning methods applying to content-based spamming detection already get good results. Link-based detection, while applying very complicated graph algorithm, the results are not bad, even not as good as the contend-based. And still, the link-based method focuses on a specified problem such as link farm or so. Still the link-based method has space to be improved.

2 Our work

There are two basic ideas in link-based spamming. One is that, many web pages with rich content were manufactured; and then, hyperlinks are hidden in those pages pointing to a spamming page. Honey pot, infiltrate, .etc methods are based on this idea. Another is make lots of web pages link to each other, so as to increase the in-link and out-link number of every page. Link exchange and link farm are based on it. The two ideas take good use of PageRank and HITS's disadvantage: the text relationship is ignored while analyzing link relationship.

Against the spamming ideas, we can make full use of the result of content-based detection method to do analyzing. A novel method proposed in this paper: cut off the links those seemed to be spamming, and then re-count the PageRank value on the re-built web network. The balance of one page's PageRank value will be regards as its probability of link-based spamming.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
ICTNET at Web Track 2010 Spam Task	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Chinese Academy of Sciences,Institute of Computing Technology,Beijing 100190,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

Presented at the Nineteenth Text REtrieval Conference (TREC 2010) held in Gaithersburg, Maryland on 16-19 November 2010. The conference was co-sponsored by the National Institute of Standards and Technology (NIST), the Defense Advanced Research Projects Agency (DARPA), and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT

Web Spamming refers those web pages deceive search engines so as to get a higher rank in their search result. We work on the data set TrecWeb09, based on a content-based spamming classifier, to check the two ends of a hyperlink; if the two end pages either is content spamming, or both are not so good, then the hyperlink will be discarded. After all hyperlinks have been checked, PageRank value shall be re-count on the re-built web network. The balance of one page?s PageRank value will be regarded as its link spamming. Then the link spamming score and the result of content deceiving analyzer will be combined as the final estimation of one page?s spamming.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**3**	

3 Data set and experiment framework

In this paper, the experiment is done on the date set TrecWeb09. This dataset was crawled from the general Web in early 2009. It contains 1 billion web pages, a substantial fraction of which are spamming. On this data set, Gordon V. Cormack [2] gives all the pages a percentile rank value indicating how much it spams, by applying a naive Bayes classifier with some content-based features. We use a pair of page ID and the percentile rank, <p_ID, p_Spamming_Percentile>, to represent the result, which is the fundamental of our later work.

If a page provides good content and clear links, and also all linked pages are not so bad, how should it be spamming? Therefore, it is easy to find some spamming features in their content. Here, we sufficiently take advantage of the result of content-based spamming detection.

For each links, we count the balance of the two ends page's percentile, and also the sum of the two's percentile. If either the balance of the sum is beyond threshold, and then the link should be discarded. We check each link in the data set, and then the net working will be re-built.

4 Experiment

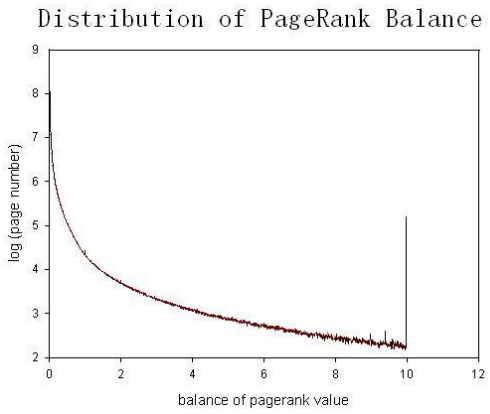
Firstly we parsed the entire original data set. Because of the huge size of it, we use Hadoop to do parsing distributed. After all pages processed, we generate the link structure of the total net working and represent it by the following format.

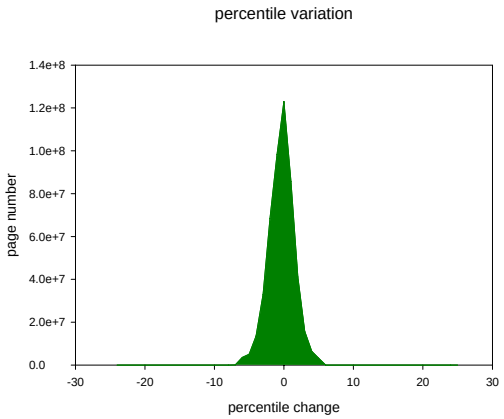
this_id INLINK_NUMBER ID1 N1 ID2 N2……

this_id means a page's id in the data set. INLINK_NUMBER refers this page's in-link page number. ID1 is the first page pointing to this page and its out-link number is N1. After the structure is represented in this format, it is easy to count the original PageRank value for each page. We use the function provided by Brin [6]

Then the net working is re-built by the above idea.

When the second PageRank value has been produced on the rebuilt networking, we get the balance of the original PageRank and the second PageRank value for each page. And further more, we regard the absolute value of the each page's PageRank balance as its score of link-based spamming.





Most balance is less than 1, totally 99.65%. It says that our algorithm didn't destruct the original web networking badly. On the other hand, the number of those pages, whose PageRank balance is more than 1, is less than 0.05% and it's far less than the proportion of spamming pages 6-8%[11].

Then we distribute all PageRank balance into the range 0-99, that is b=b*100; and the result is regarded as the punishment of link-based spamming. Then we subtract the link-based spamming punishment from the percentile given by Gordon and get a score as the spamming estimation. Finally we recount the percentile using this score.

Finally, after the recounting, 25% pages haven't changed its percentile. And all the pages' percentile variation shows in picture above.

5 Acknowledgements

We thank all the organizers of TREC Web Track and NIST. We appreciate the efforts of all assessors for judging the runs. This work is supported by NSF of China Grants No. 60873243，and also by "863" Program of China Grant No. 2008AA01Z140.

6 Reference

[1] Zoltan Gyongyi. Web Spam Taxonomy. IEEE Computer Society, 2005.

[2] Gordon V. Cormack, Mark D. Smucker, Charles L. A. Clarke, Efficient and Effective Spam Filtering and Re-ranking for Large Web Datasets, University of Waterloo, 2010

[3] Alexandros Ntoulas. Detecting Spam Web Pages through Content Analysis. WWW, 2006

[4] Luca Becchetti. Link Analysis for Web Spam Detection. ACM, 2007.

[5] Jon M Kleinberg, Authoritative Sources in a Hyperiinked Envlmnment, Jurnal of the ACM, 1999

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1–7):107–117, 1998.

[7] Xiaomeng Wu, Ichiro Ide, and Shinichi Satoh, PageRank with Text Similarity and Video Near-Duplicate Constraints for News Story Re-ranking, MMM 2010, LNCS 5916, pp. 533–544, 2010

[8] Wenpu Xing ,Ali Ghorbani. Weighted PageRank Algorithm. Proceedings of the Second Annual Conference on Communication Networks and Services search, 2004

[9] Yang Bin, Kang Muning. Concept - based Weighted PageRank Algorithm. Journal of Information, NO.11, 2006

[10] Song Liu. Pagerank on Mapreduce. Bristol University, 2010.

[11] Dennis Fetterly. Spam, Damn Spam, and Statistics, Using statistical analysis to locate spam web pages. WebDB 2004